Appl. Statist. (1999)
48, *Part* 4, *pp.* 457–468

Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis

N. E. Breslow and N. Chatterjee

University of Washington, Seattle, USA

[Received March 1998. Revised November 1998]

Summary. Two-phase stratified sampling is used to select subjects for the collection of additional data, e.g. validation data in measurement error problems. Stratification jointly by outcome and covariates, with sampling fractions chosen to achieve approximately equal numbers per stratum at the second phase of sampling, enhances efficiency compared with stratification based on the outcome or covariates alone. Nonparametric maximum likelihood may result in substantially more efficient estimates of logistic regression coefficients than weighted or pseudolikelihood procedures. Software to implement all three procedures is available. We demonstrate the practical importance of these design and analysis principles by an analysis of, and simulations based on, data from the US National Wilms Tumor Study.

Keywords: Design efficiency; Logistic regression; Nonparametric maximum likelihood; Stratified sampling

1. Introduction

Two-phase sampling was introduced by Neyman (1938) as a technique for stratification. The investigator first draws a simple random sample from the source population and classifies subjects into strata. Subsamples are drawn from each stratum and detailed covariates are measured only for individuals sampled at this second phase. By a judicious choice of strata and of within-stratum sampling ratios, such designs can yield efficient parameter estimates while minimizing the costs of the collection of data. For example, information may be available routinely for all subjects on an outcome variable (e.g. relapse) and on a mismeasured covariate. When cases of relapse and those 'positive' for the covariate are rare, it is desirable to have both categories overrepresented in the validation sample for which true covariate values are obtained (White, 1982).

This situation is well illustrated by the data in part A of Table 1 from the third and fourth clinical trials of the National Wilms Tumor Study Group (NWTSG) (D'Angio *et al.*, 1989; Green *et al.*, 1998). They show the association between treatment outcome and tumour histology for 4088 children diagnosed with the embryonal cancer of the kidney known as Wilms tumour. Patients whose tumours are composed of one of the rare cell types known collectively as 'unfavourable histology' (UH) are much more likely to relapse and die than are patients with tumours of 'favourable histology' (FH) (Beckwith and Palmer, 1978). The histologic diagnosis used for Table 1 is that of the pathologist on duty at the time of treatment at one of the more than 100 childhood cancer centres that participate in the NWTSG. The

Address for correspondence: N. E. Breslow, Department of Biostatistics, University of Washington, Seattle, WA 98195-7232, USA.
E-mail: norm@biostat.washington.edu

 0035–9254/99/48457

Table 1. Institutional histology and outcome for Wilms tumour†

Histology	*A: entire data set (N_{ij})*		*B: case–control sample (n_{ij})*		*C: balanced sample (n_{ij})*	
	Cases	*Controls*	*Cases*	*Controls*	*Cases*	*Controls*
Favourable	415	3262	415	536	415	316
Unfavourable	156	255	156	35	156	255
Total	571	3517	571	571	571	571
Odds ratio	4.8		5.8		0.47	

†Cases — relapsed; controls — not relapsed.

Table 2. Results of the weighted likelihood analyses

Variable	*Results for the following analyses:*				
	Entire data set	*2 strata*		*8 strata*	
		Case–control	*Balanced*	*Case–control*	*Balanced*
Regression coefficients					
Intercept	−2.71	−2.72	−2.57	−2.71	−2.72
Stage II	0.77	0.79	0.55	0.78	0.78
Stage III	0.77	0.69	0.48	0.79	0.81
Stage IV	1.05	1.38	1.00	1.07	1.07
UH	1.31	1.55	1.35	1.37	1.46
Stage II × UH	0.15	−0.27	0.12	0.03	−0.05
Stage III × UH	0.59	0.12	0.51	0.41	0.28
Stage IV × UH	1.26	1.02	0.98	1.01	0.91
Standard errors					
Intercept	0.11	0.12	0.13	0.11	0.11
Stage II	0.15	0.18	0.20	0.15	0.15
Stage III	0.15	0.18	0.20	0.15	0.15
Stage IV	0.18	0.24	0.26	0.18	0.18
UH	0.25	0.36	0.31	0.32	0.32
Stage II × UH	0.33	0.50	0.44	0.42	0.43
Stage III × UH	0.32	0.49	0.42	0.42	0.41
Stage IV × UH	0.39	0.83	0.62	0.60	0.63

definitive histologic diagnosis is made at the NWTSG Pathology Center by the individual pathologist who initially described the favourable and unfavourable subtypes. Compared with his readings, institutional pathologists misclassify as UH about 2% of the FH tumours and as FH nearly 30% of the UH tumours. When analysed by using logistic regression, institutional histology has no prognostic value once account has been taken of central histology. Thus institutional histology may be regarded as an error prone *surrogate* for central histology.

Apart from histology, the most important predictor of failure of treatment is the stage of disease classified as I, localized to the kidney and completely resected, II, spread beyond the kidney but completely resected, III, residual tumour in the abdomen or tumour in the lymph nodes, and IV, metastatic to the lung or liver. The second column of Table 2 shows the effects of stage and (central) histology and their interactions as modelled in a logistic regression equation using data for all 4088 subjects. The outlook is particularly bleak for children with metastatic UH disease, who fortunately comprise only 1.7% of the total.

These results required the NWTSG pathologist to examine microscopically several slides from each of 4088 tumours. Through the appropriate use of stratified sampling we aim to reduce drastically the use of central pathology. A simple random validation sample of 10% or 25% of the 4088 patients would include relatively few of the 571 relapsed cases. Much more efficient is a case–control sample (Table 1, part B) consisting of all relapsed cases plus a random sample of non-relapsed controls (Breslow, 1996). Standard computer programs may be used with such samples to estimate logistic regression coefficients (odds ratios) other than the intercept (Prentice and Pyke, 1979). However, the case–control sample contains relatively few of the rare but informative UH controls. As shown later, more efficient estimates are obtained with a 'balanced' sample (Table 1, part C) of the same size. This contains all relapsed cases, all (institutional) UH patients and about 10% of the remainder. Since the crude odds ratio associating relapse and histology in the balanced sample is less than 1, however, it is clear that some adjustments to a simple logistic regression analysis are needed to account for the biased sampling.

The essential features of this problem are the binary outcome variable, assumed known without error for all the study subjects, and the stratification used for the selection of the phase two sample. Covariates measured at phase two may be discrete or continuous. The goal is to design and analyse data from the phase two sample to approximate as accurately as possible the results that would have been obtained by fitting a logistic regression model with these same covariates to everyone. This paper illustrates methods recently developed and discussed by Scott and Wild (1997) and Breslow and Holubkov (1997a) by application to the case–control and balanced samples of NWTSG data. These methods are implemented in S-PLUS functions (MathSoft, 1996) which, together with the NWTSG data sets, are available from the authors or from Statlib (`http://lib.stat.cmu.edu`). Simulation studies demonstrate the efficiency advantages that result from a careful selection of both the phase two sample and the method of analysis.

2. Methodology

A formal description of the two-phase study is as follows. At phase one a random sample of N subjects is drawn from an infinite source population, sometimes known as a superpopulation. All subjects are classified according to a binary outcome variable Y and a stratum indicator S. Denote by N_{ij} the number with $Y = i$ and $S = j$, where $i = 0, 1$ and $j = 1, \ldots, J$. At phase two n_{ij} subjects are selected at random from among the N_{ij} that are available in each of the resultant $2J$ categories and values x_{ijk} of a p-dimensional covariate vector are measured ($k = 1, \ldots, n_{ij}$). We assume that the association between the outcome and covariates in the source population is described by the logistic regression model

$$\Pr(Y = i|X = x) = \Pr(Y = i|S = j, X = x) = \frac{\exp(\mathrm{i}x^{\mathrm{T}}\beta)}{1 + \exp(x^{\mathrm{T}}\beta)}, \tag{1}$$

where x incorporates an intercept with coefficient β_0. For problems in which S represents a discrete, error prone version of X, the assumed conditional independence of Y and S given X implies that S is a *surrogate*. For other problems, in which S is a discrete explanatory variable in its own right, we assume that its qualitative or quantitative effects are already included among the covariates X. The goal is the efficient estimation of the regression coefficients β by using both phase one $\{N_{ij}\}$ and phase two $\{x_{ijk}\}$ data.

2.1. Weighted likelihood

Three estimation methods are currently available. The first is a weighted likelihood (WL) approach with origins in sampling theory (Flanders and Greenland, 1991). If we denote by

$$U_{ijk}(\beta) = \partial[\log\{\Pr(Y = i|X = x_{ijk})\}]/\partial\beta$$

the standard logistic regression scores and by $f_{ij} = n_{ij}/N_{ij}$ the sampling fractions, $\hat{\beta}_{\mathrm{WL}}$ solves the inverse probability-weighted estimate of the score equations that would be used if covariates were available for all N subjects at phase one,

$$\hat{U}(\beta) = \sum_i \sum_j f_{ij}^{-1} \sum_k U_{ijk}(\beta) = 0. \tag{2}$$

$\hat{\beta}_{\mathrm{WL}}$ is thus known as the Horwitz–Thompson estimate. It follows from standard sampling and estimating equation theory that, under suitable regularity conditions, $\hat{\beta}_{\mathrm{WL}}$ is consistent and asymptotically normally distributed with covariance matrix that may be estimated by the 'sandwich'

$$\left(\frac{\partial \hat{U}}{\partial \beta^{\mathrm{T}}}\right)^{-1} \sum_{i,j} f_{ij}^{-2} \left\{\sum_k U_{ijk}^{\otimes 2} - \frac{1 - f_{ij}}{n_{ij}} \left(\sum_k U_{ijk}\right)^{\otimes 2}\right\} \left(\frac{\partial \hat{U}}{\partial \beta}\right)^{-1} \Bigg|_{\beta=\hat{\beta}_{\mathrm{WL}}}, \tag{3}$$

where $u^{\otimes 2}$ denotes uu^{T}. It is easily obtained by fitting a standard logistic regression model to the phase two data, using as prior weights a variable that takes values f_{ij}^{-1} for observations in the (i, j) cell.

2.2. Pseudolikelihood

Pseudolikelihood (PL) involves the maximization of a product of biased sampling probabilities that are defined as follows. Let P_{ij} and δ_j be given by

$$P_{ij} = \frac{\exp(\mathrm{i}\delta_j)}{1 + \exp(\delta_j)} = \Pr(Y = i|S = j), \tag{4}$$

so that δ_j denotes the log-odds for response $(Y = 1)$ in stratum j. Let p_{ijk} denote the probability that $Y = i$ for a subject with covariates x_{ijk} given that $S = j$ *and that the subject was sampled at phase two*. We calculate

$$p_{ijk} = \frac{n_{ij} \exp\{\mathrm{i}(\beta_0 - \delta_j + x_{ijk}^{\mathrm{T}}\beta)\}}{n_{0j} + n_{1j} \exp(\beta_0 - \delta_j + x_{ijk}^{\mathrm{T}}\beta)}. \tag{5}$$

The PL estimate of Schill *et al.* (1993) maximizes the PL

$$L_1 L_2 = \prod_{i,j} P_{ij}^{N_{ij}} \prod_{i,j,k} p_{ijk}$$

as a function of the $(J + p)$-dimensional parameter vector $\gamma = (\delta^{\mathrm{T}}, \beta^{\mathrm{T}})^{\mathrm{T}}$. Breslow and Holubkov (1997b) described in detail how this may be accomplished by fitting a logistic regression model, with appropriately defined offset and design matrix, jointly to the phase one and two data. The usual covariance matrix is adjusted by subtraction of an appropriate correction term.

A slightly simpler PL estimate is obtained by maximizing L_1 alone to obtain $\hat{\delta}_j = \log(N_{1j}/N_{0j})$ and substituting in L_2 to find $\hat{\beta}_{\mathrm{PL}}$. This is accomplished by fitting the logistic regression model (1) to the phase two data using an offset with values $\log(n_{1j}N_{0j}/n_{0j}N_{1j})$ for subjects in stratum

j to correct for the biased sampling (Breslow and Cain, 1988). These two versions of the PL estimate yield very similar results in practice.

2.3. Nonparametric maximum likelihood

Since the marginal distribution of the stratum and covariates (S, X) has been left unspecified, equation (1) defines a problem in semiparametric inference. For fixed β, the nonparametric maximum likelihood (NPML) estimate of the marginal distribution places mass on the observed values x_{ijk} of X (Gill *et al.*, 1988). Hence NPML estimation for the semiparametric problem corresponds to ordinary ML estimation for the problem where X is discrete, albeit taking a large number of values. Scott and Wild (1991, 1997) solved this problem for simple random sampling at phase one; Breslow and Holubkov (1997a) solved it for case–control sampling at phase one (see the next section). Although obtained by using different approaches, the two solutions are identical and may be described as follows. Define $\xi_j = \xi_j(\delta_j)$ by

$$\xi_j = \log\left(\frac{n_{1j} - N_{1j} + N_{+j}P_{1j}}{n_{0j} - N_{0j} + N_{+j}P_{0j}}\right) - \delta_j$$

and $\tilde{p}_{ijk} = \tilde{p}_{ijk}(\delta_j, \beta)$ by

$$\tilde{p}_{ijk} = \frac{\exp\{\mathrm{i}(\xi_j + x^{\mathrm{T}}\beta)\}}{1 + \exp(\xi_j + x^{\mathrm{T}}\beta)}.$$

The NPML estimate $\hat{\gamma} = (\hat{\delta}^{\mathrm{T}}, \hat{\beta}^{\mathrm{T}}_{\mathrm{ML}})$ solves the J equations ($j = 1, \ldots, J$)

$$N_{1j} - N_{+j}\,P_{1j}(\delta_j) = \sum_i \sum_k \epsilon_i\{1 - \tilde{p}_{ijk}(\delta_j, \beta)\} \tag{6}$$

and the p-dimensional equation

$$\sum_i \sum_j \sum_k \epsilon_i\{1 - \tilde{p}_{ijk}(\delta_j, \beta)\}x_{ijk} = 0, \tag{7}$$

where $\epsilon_1 = -\epsilon_0 = 1$. A logistic regression program again suffices for fitting, but this time it requires iterative application using a Gauss–Seidel approach (e.g. Jacquez (1970), p. 171), as follows. Starting with $\hat{\delta}_j = \log(N_{1j}/N_{0j})$ and solving equation (7) for β yields the Breslow and Cain (1988) version of $\hat{\beta}_{\mathrm{PL}}$. Fitted values $\tilde{p}_{ijk}$ are inserted into the right-hand side of equation (6) which is then solved for δ_j, or equivalently ξ_j, and the process is repeated. This algorithm may be slow or fail to converge. Therefore the joint solution of equations (6) and (7) by using standard numerical techniques is often preferable. It is important to start the iteration at $\hat{\beta}_{\mathrm{PL}}$ and to search for local roots since multiple solutions may exist. Asymptotic covariance matrices for $\hat{\beta}_{\mathrm{ML}}$ were given by Scott and Wild (1997) and by Breslow and Holubkov (1997a, b).

2.4. Outcome-dependent sampling at phase one

A slightly more complex sampling design involves separate samples of N_1 cases ($Y = 1$) and N_0 controls ($Y = 0$) drawn from the source population at phase one. Without additional information it is then impossible to estimate consistently the intercept β_0 in equation (1), even if complete covariate data are available for all $N_0 + N_1 = N$ subjects. We therefore assume for the sake of argument that

$$\alpha = \log\{\Pr(Y = 1)/\Pr(Y = 0)\},$$

the marginal log-odds of response, is known. With some redefinition of parameters, and adjustment of covariance matrices to account for the knowledge of α, the preceding methodology still applies. Fortunately the covariance adjustments and the assumed value for α only affect δ and β_0, not the regression coefficients of primary interest. Thus the fact that α may be unknown is of little consequence.

For this new biased sampling scheme we redefine P_{ij} in equation (4) to equal the probability P_{ij}^* that $Y = i$ given that $S = j$ *and that the subject was sampled at phase one*. If δ_j^* denotes the log-odds of response in stratum j under this new scheme, then

$$P_{ij}^* = \frac{\exp(\mathrm{i}\delta_j^*)}{1+\exp(\delta_j^*)} = \frac{N_i \exp\{\mathrm{i}(\delta_j - \alpha)\}}{N_0 + N_i \exp(\delta_j - \alpha)}.$$

Estimation in the reparameterized model using equations (6) and (7) yields estimates $\hat{\delta}_j^*$ and $\hat{\beta}^*$ to which parameters in the original model are related via

$$\hat{\delta}_j = \hat{\delta}_j^* + \alpha - \log(N_1/N_0), \qquad j = 1, \ldots, J,$$

$$\hat{\beta}_0 = \hat{\beta}_0^* + \alpha - \log(N_1/N_0)$$

and $\hat{\beta}_l = \hat{\beta}_l^*$, $l = 1, \ldots, p-1$. The fact that the distribution of the phase one data is changed from a single multinomial $\{N_{ij}\}$ to a pair of independent multinomials $\{N_{0j}\}$ and $\{N_{1j}\}$ leads to a reduction in the asymptotic variance of the estimating functions (scores) and hence to a reduction in the asymptotic variance of the parameter estimates. This reflects the additional information contributed by the assumption that α is known. For the WL estimate the reduction is achieved by subtraction of

$$\sum_i N_i^{-1} \left(\sum_j f_{ij}^{-1} \sum_k U_{ijk} \right)^{\otimes 2}$$

from the middle term in equation (3). For the PL and ML estimates, the variance of β_0 is reduced by $1/N_0 + 1/N_1$; see Holubkov (1995).

3. Results and simulations

The results of fitting the interaction model to the case–control and balanced phase two samples of NWTSG data by using WL are shown in the third and fourth columns of Table 2. Corresponding results for ML are presented in Table 3. By comparing the regression coefficients and standard errors with those for the complete data set (second column), we draw the following tentative conclusions:

(a) there is little difference between WL and ML for the case–control design;
(b) the balanced design is superior for the estimation of interactions for both WL and ML, whereas the case–control design is slightly better for the estimation of the main effects of stage;
(c) ML is better than WL for the balanced design.

Results for PL (not shown) were intermediate.

The preceding analysis used just two strata for sampling cases and controls, namely histology (FH or UH) as evaluated by the institutional pathologist. This wastes information since stage was also determined by the institution and hence was known for all patients. A

Table 3. Results of the ML analyses

Variable	*Results for the following analyses:*				
	Entire data set	*2 strata*		*8 strata*	
		Case–control	*Balanced*	*Case–control*	*Balanced*
Regression coefficients					
Intercept	−2.71	−2.72	−2.57	−2.71	−2.71
Stage II	0.77	0.80	0.55	0.77	0.77
Stage III	0.77	0.69	0.49	0.78	0.79
Stage IV	1.05	1.38	0.97	1.07	1.05
UH	1.31	1.57	1.26	1.36	1.38
Stage II × UH	0.15	−0.25	0.29	0.09	0.10
Stage III × UH	0.59	0.13	0.73	0.49	0.46
Stage IV × UH	1.26	1.07	1.43	0.98	1.37
Standard errors					
Intercept	0.11	0.12	0.13	0.11	0.11
Stage II	0.15	0.18	0.20	0.15	0.15
Stage III	0.15	0.18	0.20	0.15	0.15
Stage IV	0.18	0.24	0.25	0.18	0.17
UH	0.25	0.36	0.29	0.30	0.28
Stage II × UH	0.33	0.51	0.39	0.39	0.37
Stage III × UH	0.32	0.49	0.37	0.39	0.35
Stage IV × UH	0.39	0.85	0.47	0.55	0.44

second set of analyses, for which results are presented in the fifth and sixth columns of Tables 2 and 3, used eight strata formed by the cross-classification of institutional histology and stage. The original results for the main effects of stage are now reproduced almost exactly by both WL and ML. For balanced sampling there is little or no improvement in the WL standard errors for the UH and interaction effects compared with those obtained earlier. ML seems better able to utilize the additional phase one data given by the finer stratification, especially for the case–control design.

We conducted a simulation study to determine whether these general conclusions would hold once the vagaries of (phase two) sampling were eliminated. 100 separate phase two samples were drawn for each design and logistic regression models were fitted to them by using WL, ML and PL. Figs 1 and 2 show graphically the mean-squared error (MSE) of each regression coefficient, with the original coefficients (entire data set) considered as the true values. This confirms the essential features already noted from the illustrative analysis. WL and ML are virtually equivalent for the case–control design when only two strata are used at phase one. The MSE is reduced much more under ML than under WL by incorporating stage at phase one or by using the balanced design. Results for PL are intermediate; they generally agree with those of WL for the case–control design and with ML for the balanced design.

Reilly and Pepe (1995) determined optimal phase two sampling fractions for a fixed total phase two sample size by studying the asymptotic variances of regression coefficients estimated by WL. Table 4 contrasts the optimal designs for estimating each of the three interaction terms with the case–control and balanced designs, treating the entire data set as the population. The optimal designs sample all the especially rare and informative UH cases but represent a compromise between the case–control and balanced designs in sampling the UH controls.

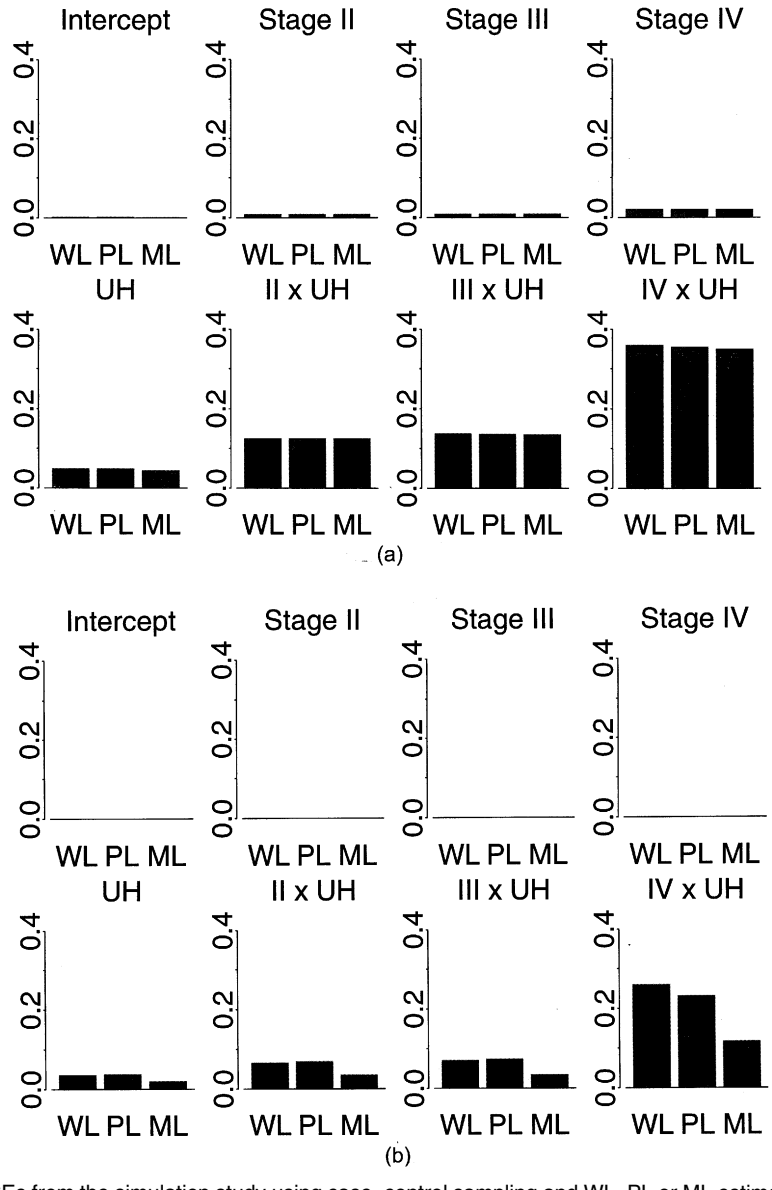

Fig. 1. MSEs from the simulation study using case–control sampling and WL, PL or ML estimation: (a) stratified by outcome and histology; (b) stratified by outcome, histology and stage

Fig. 3 graphs the average standard errors observed for the interaction terms in 100 phase two samples drawn according to the case–control, balanced and (corresponding) optimal designs. There is little to choose between the balanced and optimal designs, both of which are superior to the case–control design. Although imperceptible to the eye, the average standard errors for PL and WL for the balanced design are in some cases actually less than those for the optimal (WL) design. For example, for estimation of the stage IV $\times$ UH interaction, the average standard error for WL was 0.549 for the optimal and 0.551 for the balanced design. By contrast, the corresponding average standard errors for ML were 0.497 and 0.472.

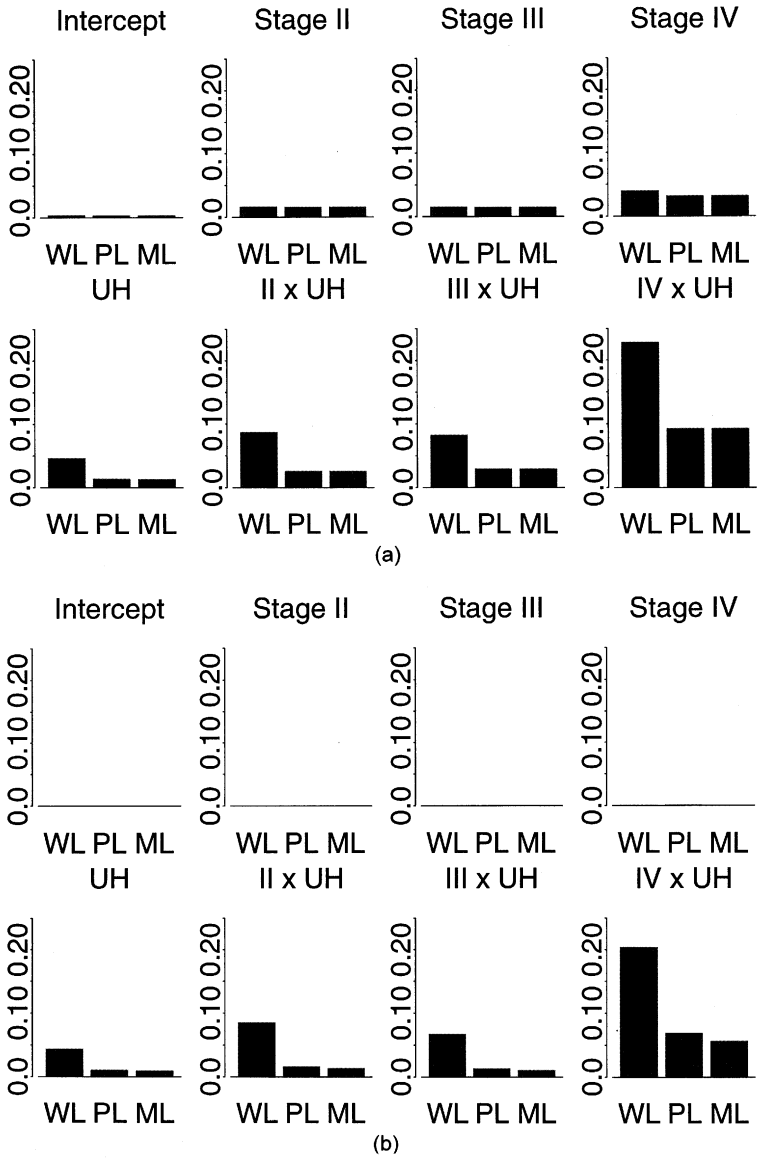

Fig. 2. MSEs from the simulation study using balanced sampling and WL, PL or ML estimation: (a) stratified by outcome and histology; (b) stratified by outcome, histology and stage

4. Discussion and conclusions

A comparison of the second and last columns of Table 3, or an examination of the ML results for the second part of Fig. 2, demonstrates very close concordance between logistic regression coefficients estimated by using the entire data set of 4088 records and those obtained by using ML for the balanced, finely stratified sample. Yet the latter required central histology evaluation for only 1142 patients (28%). Since institutional histology and stage are required for the determination of the treatment, this means that the essential results of this study could

Table 4. Phase two sampling fractions (n_{ij}/N_{ij})

Design	*Fractions for the following groups:*			
	Cases		*Controls*	
	FH	*UH*	*FH*	*UH*
Case–control	1.00	1.00	0.16	0.16
Balanced	1.00	1.00	0.10	1.00
Optimal				
Stage II × UH	0.82	1.00	0.15	0.65
Stage III × UH	0.83	1.00	0.14	0.72
Stage IV × UH	0.62	1.00	0.17	0.75

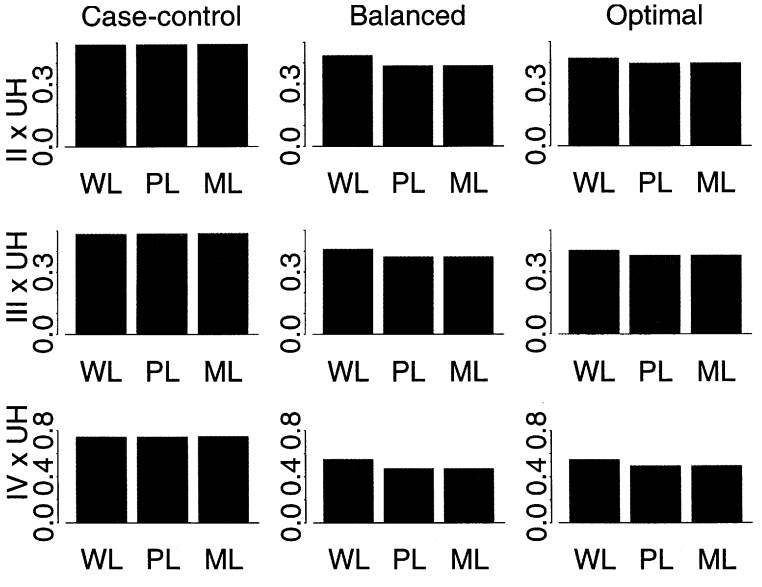

Fig. 3. Average standard errors for interaction coefficients from the simulation study using WL, PL or ML estimation, stratified by outcome and histology

have been obtained with sharply lower (marginal) costs by a careful selection of the slides sent to the NWTSG pathologist. This of course presumes that slides will be stored for the 2 or 3 years that are needed to determine the outcome and that central pathology is not required for other purposes.

Three factors contributed to this desirable result. First, even though phase two sampling was based only on institutional histology and outcome, the strata incorporated institutional stage as well. This illustrates the basic principle that one should incorporate at phase one as much of the available data as possible. The obvious limitation is that the phase one data are necessarily discrete. With smaller sample sizes, too fine a stratification may leave some cells empty, leading to a breakdown in the methodology due to infinite or indeterminate values for some sampling fractions or $\hat{\delta}_j$.

The second contributing factor was the use of a reasonably informative design, in this case the 'balanced' design. Balance was defined in an *ad hoc* fashion here to mean sampling of all relapsed cases and all those with UH tumours. Elsewhere (Breslow and Cain, 1988) it has

been defined more precisely to mean the selection of $\{n_{ij}\}$ that are as equal as possible given the available $\{N_{ij}\}$ and the total phase two sample size. Even greater overall efficiency would result if we had applied this principle to the *eight* strata formed by histology and stage, to ensure that the phase two sample included more of the 330 available FH stage IV controls than it did. The attempt to achieve rough parity in numbers of phase two subjects in each of the $2J$ cells through the use of a 'balanced' design does not always achieve near optimality, especially not when the model is restrictive. For quantitative linear regression, for example, it is better to sample at the extremes of the covariate space. In many practical situations, however, the balanced design represents a reasonable compromise between the competing demands of efficiency and the need to check model assumptions of linearity or additivity (Breslow and Cain, 1988). Also, as demonstrated here, optimality of the design depends on the particular method of estimation; the most efficient design using WL is not necessarily the most efficient for ML. Reilly (1996) has developed optimal designs for WL estimation when the total cost of observation, with different costs at phases one and two, is fixed.

The third factor that contributed to the close concordance of estimates was the use of an efficient estimation method, namely ML. This would be even more important with the less efficient case–control design (Fig. 1). Apart from the increased computational complexity, the primary drawback of ML comes about when we are attempting to fit the *wrong model*. For example, for administrative purposes we may want to know the slope of the best fitting (logistic) linear regression model for the population, even in the presence of some curvature. The great advantage of the Horwitz–Thompson (WL) estimate in such circumstances is that it will consistently approximate the result that would be obtained if covariate values were available for all subjects. By contrast, ML and PL do not (Scott and Wild, 1986; Xie and Manski, 1989). Thus ML achieves efficiency at the cost of a certain lack of robustness. This was not an issue for the NWTSG study since we used a saturated (interaction) model. It could be if an additive model were fitted instead.

Two-phase sampling, as considered here, is perhaps the simplest example where data are missing by design. Whittemore (1997) developed WL estimation for sampling conducted in three or more phases involving nested partitions of the sample space into increasingly fine strata, exploiting the fact that the resulting data are subject to a monotone pattern of missingness. Besides the fact that it is broadly applicable to such problems, WL estimation enjoys the advantages of being relatively easy to implement and, as just mentioned, robust to model misspecification. Wacholder *et al.* (1994) developed PL estimation for their partial questionnaire design, whereby different subsets of subjects are missing different (non-nested) sets of covariate values and where complete data are available for one subset. Since their procedure requires the consistent estimation of the joint covariate distribution, it is currently restricted to situations where there are a small number of discrete covariates. Fully efficient ML or NPML estimation methods have yet to be developed for these more complex designs.

Acknowledgements

This work was supported in part by US Public Health Service grant CA40644. Helpful comments by the Joint Editor and referees are acknowledged with gratitude.

References

Beckwith, J. B. and Palmer, N. F. (1978) Histopathology and prognosis of Wilms tumor. *Cancer*, **41**, 1937–1948.
Breslow, N. E. (1996) Statistics in epidemiology: the case-control study. *J. Am. Statist. Ass.*, **91**, 14–28.

Breslow, N. E. and Cain, K. C. (1988) Logistic regression for two-stage case-control data. *Biometrika*, **75**, 11–20.
Breslow, N. E. and Holubkov, R. (1997a) Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *J. R. Statist. Soc.* B, **59**, 447–461.
———(1997b) Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data. *Statist. Med.*, **16**, 103–116.
D'Angio, G. J., Breslow, N., Beckwith, B., Evans, A., Baum, E., Delorimier, A., Fernbach, D., Hrabovsky, E., Jones, B., Kelalis, P., Othersen, B., Teft, M. and Thomas, P. R. M. (1989) Treatment of Wilms' tumor. *Cancer*, **64**, 349–360.
Flanders, W. D. and Greenland, S. (1991) Analytic methods for two-stage case-control studies and other stratified designs. *Statist. Med.*, **10**, 739–747.
Gill, R. D., Vardi, Y. and Wellner, J. A. (1988) Large sample theory of empirical distribution in biased sampling models. *Ann. Statist.*, **16**, 1069–1112.
Green, D. M., Breslow, N. E., Beckwith, J. B., Finklestein, J. Z., Grundy, P. G., Thomas, P. R. M., Kim, T., Shochat, S., Haase, G. M., Ritchey, M. L., Kelalis, P. P. and D'Angio, G. J. (1998) Comparison between single-dose and divided-dose administration of dactinomycin and doxorubicin for patients with Wilms tumor: a report from the National Wilms Tumor Study Group. *J. Clin. Oncol.*, **16**, 237–245.
Holubkov, R. (1995) Maximum likelihood estimation in two stage case-control studies. *PhD Dissertation*. University of Washington, Seattle.
Jacquez, J. A. (1970) *A First Course in Computing and Numerical Methods*. Reading: Addison-Wesley.
MathSoft (1996) *S-Plus*. Seattle: MathSoft.
Neyman, J. (1938) Contribution to the theory of sampling from human populations. *J. Am. Statist. Ass.*, **33**, 101–116.
Prentice, R. L. and Pyke, R. (1979) Logistic disease incidence models and case-control studies. *Biometrika*, **66**, 403–411.
Reilly, M. (1996) Optimal sampling strategies for two phase studies. *Am. J. Epidem.*, **143**, 92–100.
Reilly, M. and Pepe, M. S. (1995) A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, **82**, 299–314.
Schill, W., Jöckel, K. H., Drescher, K. and Timm, J. (1993) Logistic analysis in case-control studies under validation sampling. *Biometrika*, **80**, 339–352.
Scott, A. J. and Wild, C. J. (1986) Fitting logistic models under case-control or choice based sampling. *J. R. Statist. Soc.* B, **48**, 170–182.
———(1991) Fitting logistic regression models in stratified case-control studies. *Biometrics*, **47**, 497–510.
———(1997) Fitting regression models to case-control data by maximum likelihood. *Biometrika*, **84**, 57–71.
Wacholder, S., Carroll, R. J., Pee, D. and Gail, M. H. (1994) The partial questionnaire design for case-control studies. *Statist. Med.*, **113**, 629–634.
White, J. E. (1982) A two stage design for the study of the relationship between a rare exposure and a rare disease. *Am. J. Epidem.*, **115**, 119–128.
Whittemore, A. S. (1997) Multistage sampling designs and estimating equations. *J. R. Statist. Soc.* B, **59**, 589–602.
Xie, Y. and Manski, C. F. (1989) The logit model and response-based samples. *Sociol. Meth. Res.*, **17**, 283–302.